RESEARCH PAPER

A Semantic Approach to Workflow Management and Reuse for Research Problem Solving

Nikolay A. Skvortsov[1†] & Sergey A. Stupnikov

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow 119333, Russia

Keywords: Workflow reuse; Workflow patterns, Domain ontology; Canonical workflow framework for research, CWFR; Principles of FAIR data

Citation: Skvortsov, N.A., & Stupnikov, S.A.: A semantic approach to workflow management and reuse for research problem solving. Data Intelligence 4(2), 439-454 (2022). doi: 10.1162/dint_a_00142
Received: September 20, 2021; Revised: December 29, 2021; Accepted: February 4, 2022

ABSTRACT

The investigation proposes the application of an ontological semantic approach to describing workflow control patterns, research workflow step patterns, and the meaning of the workflows in terms of domain knowledge. The approach can provide wide opportunities for semantic refinement, reuse, and composition of workflows. Automatic reasoning allows verifying those compositions and implementations and provides machine-actionable workflow manipulation and problem-solving using workflows. The described approach can take into account the implementation of workflows in different workflow management systems, the organization of workflows collections in data infrastructures and the search for them, the semantic approach to the selection of workflows and resources in the research domain, the creation of research step patterns and their implementation reusing fragments of existing workflows, the possibility of automation of problem-solving based on the reuse of workflows. The application of the approach to CWFR conceptions is proposed.

1. INTRODUCTION

Providing the reuse of data includes organizing the processes of their processing and analysis by both humans and machines. Workflows are applied to describe complex processes and control their recurring execution for data processing and analysis or instance in research problem-solving or experimenting. The same workflows can be used to solve similar problems in different situations by different researchers in

† Corresponding author: Nikolay A. Skvortsov (Email: nskv@mail.ru; ORCID: 0000-0003-3207-4955).

communities and on the same or different data. Thus, once developed, a workflow can be reused multiple times, thereby serving the reuse of research data and reproducing research results.

The guiding principles of FAIR data [1] serve to provide the reuse of data and may be supported by accompanying the data with workflows suitable for their processing. In this case, the FAIR data principles should be fully applied to processed data and related workflows as a kind of data. To support the FAIR properties of data, it is necessary that all metadata and resources referred to the data during their lifecycle and applied to them comply with the principles of FAIR data. After all, the FAIR data guidelines are equally guiding for managing research workflows since they are a type of data, a type of metadata, and in their turn require being described by the metadata. This means that workflows should be findable (F), accessible (A), interoperable (I), and reusable (R).

To manage the lifecycle of research problem solving and ensure the reuse and reproducibility of research results, it is important to provide the search for relevant workflows, their interoperability in data infrastructures, and correct implementations of research processes.

The conceptions of the Canonical Workflow Framework for Research (CWFR) [2,3] include the creation of canonical steps of activities for typical research approaches common to most domains. Research patterns consist of canonical steps. The step descriptions are available from the libraries of canonical steps. For different contexts, domains, and communities, there are libraries of specialized packages for certain step patterns.

The interoperability of workflows is supported by using FAIR digital objects [4]. During the execution of each step of the workflow, a digital object is formed, labeled with a unique persistent identifier, defined by the type, described by attributes, and having a state contributed by the completed steps. Access to any state is possible via digital object identifiers. The use of canonical workflow step patterns for solving research problems allows to standardize the research process, ensure the necessary formal steps, and simplify the reuse of workflows and processed data. In addition, workflow step patterns and typing the steps in specific domains are a good hint to the machine on how to process data.

The principles of FAIR data initially declare machine-actionability, this property should be applied to workflows as well. According to this principle, the reuse of workflows does not mean that once developed by a human, the workflow can be reused by a machine to solve the same problem. But this means that a machine that had not yet worked on the problem should be able to semantically analyze the problem, find relevant data, find the way to obtain the necessary result, possibly by creating a new workflow for solving the problem. So, it can solve it and publish both the obtained results and the created tools in such a way that both humans and machines can find and apply them for further reuse or to reproduce the results.

Semantic approaches for these purposes can be based on the use of the domain and special ontologies. The role of ontologies in semantic approaches to describing workflows was emphasized in works related to the myExperiment research community [5] and later Research Objects (RO) [6]. The last one is popular

and useful today. The ideas used in this study are the development and translation of the proposals reflected in the work [7].

The myExperiment project was popular for some communities since there were libraries of domain-specific services and workflows findable by keywords. The weakness was that most of the available workflows shared by researchers did not implement domain methods but just access to specific data resources. The project was still useful but faded in popularity because other tools like Python libraries were more targeted on how to solve problems in various domains. Jupyter Notebook became popular as well because Python was popular, not since it is so good for sharing and reusing resources. Similarly, GitHub is not a comfortable research infrastructure, but it stores lots of parametrized and implemented methods findable through basic search engines. These instruments have serious limitations and manual approaches from the view of data interoperability and integration. Thus, semantically searchable collections of workflows oriented to the implementation of research methods and their quality classification in domains could be useful for communities and for organizing their activities and relationships.

This concept paper related to advanced workflow technologies proposes extending investigations in CWFR with an ontological approach to the definition of workflow semantics from points of view of different ontologies and different levels of workflow definitions. Three levels of workflow semantics definition are used here. The ontological level describes the semantics of workflows, activities, and their elements in terms of domain concepts. The data model layer defines the control patterns used in different languages and workflow management systems (for example, BPMN, YAWL). At the workflow specification level, data transformation specifications and the semantics of research data processed by workflows are defined. The ontology-based semantic approach to workflow management provides a semantic search for workflows for their reuse and ensures interoperability at all levels. The following sections describe levels of workflow description and the way in which these descriptions can be applied for workflow reuse and interoperability. Then an example of workflow step patterns for problem-solving and modeling are described.

2. THE PRINCIPLES OF SEMANTIC DESCRIPTION OF WORKFLOWS

For semantic annotation of workflows, domain ontologies and special ontologies for different purposes are used simultaneously. Domain ontologies are necessary to determine the semantics of workflows, activities, and places (or inputs and outputs) from the point of view of the research problem being solved. They are also used to simplify the possibility to reuse them in the domain. Ontologies related to the research lifecycle define concepts of standard or preferred procedures for achieving research objectives. They can include such processes as the stages from problem statement to reporting on their solution, the stages of transformation and integration of heterogeneous data, the stages of testing hypotheses and theories following the scientific method, the stages of applying machine and deep learning, verification of decisions being made, and other special approaches to research. The data provenance [8] ontology defines authorship, licensing, relevance, provenance, and other non-functional information about workflows and data. Semantic annotation can be defined as expressions of concepts of several ontologies to define an annotated object from different angles of view.

A comprehensive formal ontological description and requirements in terms of ontologies allow finding relevant workflows, their fragments, individual activities, and resources for problem-solving. Means for collecting and classifying workflows, as well as searching for them by ontological descriptions, are of fundamental importance. They are necessary to offer a way for problem-solving to experts using the available resources and research results of the domain community, and when using verifiable formal approaches, it makes possible automated decisions for the reuse of workflows.

3. DEFINITION OF WORKFLOW CONTROLLING PATTERNS

At the level of workflow data models, it is suggested to use workflow patterns (workflowpatterns.com) [9]. By control patterns, metamodels are described that define constructs providing certain control rules in workflow languages. The model and semantics of such constructs are defined and workflow languages that use them are listed. The workflow metamodel ontology defines concepts for various patterns, allows expressing and annotating workflows in their terms, and provides finding relevant activities taking into account the semantics of control patterns.

There are patterns with very different semantics. The basic control patterns are sequence, parallel split (And-Split), synchronization (And-Join), exclusive choice (Xor-Split), simple merge (Xor-Join). The synchronization control pattern (Fig. 1) specifies that several branches and joined to one when all input branches are enabled. In the ontology of workflow control patterns, the AndJoin concept can be defined as having multiple “inputBranch” relations and single “outputBranch” relation with the Activity concept.

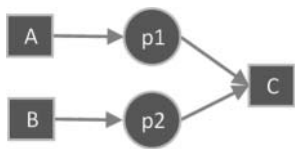


Figure 1. An example of a synchronization control pattern.

An input of the workflow activity can be annotated in terms of this ontology as an instance AndJoin with specified Activity concept (or subconcept) instances in addition to the definition of the research domain semantics of the type of this activity input and other descriptions. Specifying workflow annotations in terms of this ontology makes it possible to transfer workflows from one system to another or execute them in their systems. In any case, their use will meet the specifications.

4. DEFINITION OF DOMAIN SEMANTICS OF WORKFLOWS

Interoperability at the level of workflow specifications is provided using languages of workflow management systems directly. Annotating activities with a verbal description, simple linking with concepts, keywords, or terms from a domain dictionary is insufficient for expressing semantics readable for humans and machines. The annotation should determine not only the domain meaning of workflow elements but

link them with other elements in semantics descriptions for the possibility of analyzing the semantics of the system or its parts. The description of the semantics of data transformation and integrity requirements can be provided using preconditions and post-conditions of workflows as a whole and separate activities. Types of dataflow can be defined for inputs/outputs (places) between activities. If the expressive power of the workflow specification languages does not provide capabilities for that, the restrictions are defined only at the ontological level.

In addition to the domain related to the research object, it is desirable to define ontologies related to the most general knowledge about performing research, the requirements of the scientific method and open science, and others. Research infrastructure collections acquire workflows that are accessible using semantic search. One of the types of acquired specifications is research workflow patterns that describe required or preferred steps of workflows providing the research lifecycle or specific types of research. Research step patterns should be annotated in detail in terms of ontologies of research. The workflow patterns and their activities are backed by collections of respective implementations of workflows, services, resources, stored in digital object containers. Their annotations are more specialized or equal to the specifications of workflow step patterns. In various domains, these implementations may use specialized resources known and available in research communities.

5. SCENARIOS OF APPLYING THE SEMANTIC DESCRIPTIONS OF WORKFLOWS

During problem-solving, researchers can reuse both workflow patterns to implement them, or existing implementations of methods and processes. Semantic annotations of workflow step patterns can be queries to find their implementations. Workflow patterns, having rich metadata in terms of ontologies, are the key to creating well-annotated implementations. Formal descriptions as metadata allow applying automated reasoning for selecting and reusing relevant workflows.

Workflows can be developed with the analysis of requirement models of research problems. Depending on the problem statement type (such as data acquiring and analysis, modeling, or machine learning), different research step patterns are reused. Partly, the requirement models and workflows are developed using the found patterns. Then some fragments of the patterns can be implemented, or relevant existing workflow implementations from the collection can be found for them.

Ontological reasoning allows formal automatic search for relevant workflows or their fragments, verifying, and controlling the compliance of substituted parts of workflows and implementations on the semantic level. To find and verify resources relevant to requirements, the semantic annotation of the resource must belong to an equivalent concept or a subconcept of the concept defined by requirement annotation. So relevant research patterns can be found by the problem statement knowledge as subconcepts of requirement, or vice versa the pattern step can be proposed if the problem statement mentions its subconcepts. Implementations replacing the pattern activities should follow the requirements (to be subconcepts) imposed by the patterns and by the problem statements.

Workflow step patterns may be implemented not completely but partially if there are no requirements for full implementation. At the same time, descriptions of the precondition and post-condition requirements between workflow activities after removing a step should be reconciled using ontological reasoning and making simple transformations for correct reconnection of the fragments of patterns and their implementations.

Workflow pattern fragments can be implemented by relevant fragments of existing workflows found in registries based on logical inference. Not only the relevance of activities but also the relevance of inputs and outputs is evaluated. For the implementation of the activity, its semantics and the types of its inputs and outputs defined in semantic annotations are analyzed taking into account all the ontologies used including the data model compatibility. The search for activities with relevant types of input and activities with relevant types of output is performed in combination with the reachability problem solution between them [10]. The compliance of the semantics of the implementation with the requirements of the pattern can also be checked taking into account the role chains. An example of required substitution is given below (Fig. 2). Let the upper workflow be the canonical steps for modeling the research object. It needs substitution by some implemented workflow fragments. The lower one is a workflow used for substitution. Activity "Getting Observational Data" (let's sign it G by the first letter) can be implemented by a workflow fragment consisting of two activities "Extract Real-World Object Data" (E) and "Loading and transforming Data" (L) if the input type of activity G is a subtype of the input type of activity E (weakened precondition), the output type of G is a supertype of the output type of L (strengthened post-condition), and E is achievable from L. Semantic correspondence of activities themselves can be a non-trivial problem since the activities can use very different granularity of similar processes. Different methods can be applied for provable refining G by the fragment E-L or just evaluation if as a similarity between them.

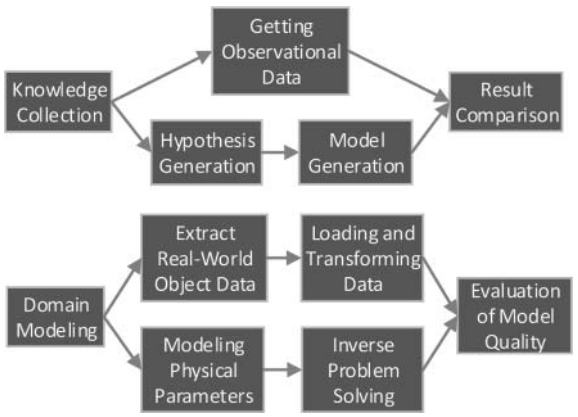


Figure 2. A workflow substitution for canonical steps of modeling.

The described approach can take into account the implementation of workflows in different workflow management systems, the organization of workflows collections in data infrastructures, and the search for them. It can be applied for the semantic approach to the selection of workflows and resources in the

research domain, the implementation research step patterns by reusing fragments of existing workflows, and automation of problem-solving based on the reuse of workflows.

6. A CANONICAL RESEARCH WORKFLOW EXAMPLE

In the frame of the investigations in research infrastructures, a lifecycle for problem-solving was developed [11]. It includes setting a problem as requirement specifications, describing requirements in terms of domain knowledge, searching for relevant data sources, integrating them, operationalizing requirements, searching or implementing methods and workflows, and experimenting. In addition, at different stages of the lifecycle, formal verification of the performed stages can be provided and the results of finished stages can be published in the domain community. The integration stage can be specified as a sequence of three data management tasks: data model (data definition and manipulation language) reconciliation, then schema-matching, and finally entity resolution in data from different sources (Fig. 3).

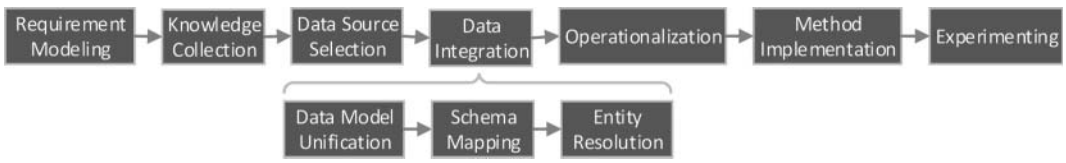


Figure 3. The data-driven problem-solving lifecycle canonical steps.

Another approach to experimenting especially in the case of investigating unobservable physical parameters of the research object is modeling it (see the canonical steps in Fig. 2). Hypotheses are generated using domain knowledge. A model is generated according to that knowledge and the hypotheses. Then distributions of modeled parameters are compared to observed ones. These recurring approaches to research on data are represented as steps of canonical workflows.

The shown examples of canonical research patterns are domain-independent. A lot of domains use similar approaches to problem-solving lifecycle beginning with gathering data from multiple sources then integrating them into the information system, identifying the same entities in their data, and applying methods to the consolidated data to solve the problem and publish the result. As well, modeling real-world objects and comparing them to observed data from them is known as a way to check research hypotheses.

In [12], the solution to astronomical problems was presented, which can be an example of the problem-solving lifecycle application with data reuse at different stages. The first described problem was finding hierarchical multiple stellar systems among data on binary stars. The requirement model of this problem was created as a decomposition tree. The domain knowledge of binary and multiple stars was accumulated in the domain ontologies and conceptual schemas for domain data representation. The ontologies and schemas were published for further reuse. Evaluating ontological relevance, heterogeneous catalogs of binary stars were integrated into the conceptual scheme, their structural heterogeneity was resolved. The results of integration were published to be used in the research community. Then, the algorithm of binary

and multiple stellar system cross-matching was performed, in which observed parameters of stars were used to identify the same stellar systems in data from different catalogs. The list of identifications was published as a new catalog [13]. Finally, the identified systems were analyzed for compliance with the hypothesis of hierarchical stellar systems (see Fig. 4). All steps of the problem solving can be implemented as implementations of the research lifecycle workflow step patterns. The ontological descriptions of these implementations comply with the descriptions of the step patterns and refine them. So the implementations can be registered using those formal descriptions and further can be found using requirements of the steps patterns in combination with some problem requirements.

Figure 4. The steps of solving the problem of finding hierarchical stellar systems.

After the problem of hierarchical stellar systems had been solved, another problem of creating a Galaxy model of binary stars began to be solved. The published results of solving the previous problem in the same domain were partially reused. Almost all stages used those results including the ontologies, the schemas, the integrated catalogs of binary stars, and the list of cross-matched binary systems. At the same time, hypotheses of the distributions of the binary stars in the Galaxy were formed using domain knowledge and relevant publications. Following the hypotheses, the Galaxy models of binary stars were generated. Then, to select the best hypotheses, the distribution of binary star parameters according to the data from catalogs were compared to the distributions of stars in the visible part of the generated Galaxy models. For more details on data reuse, see [12].

The solution to these research problems can be presented as refinements of the steps of the problem-solving canonical workflow in the domain of binary stars combined with the canonical steps of modeling. And the published results of solving the first problem could be found and reused for solving the second problem (Figure 5).

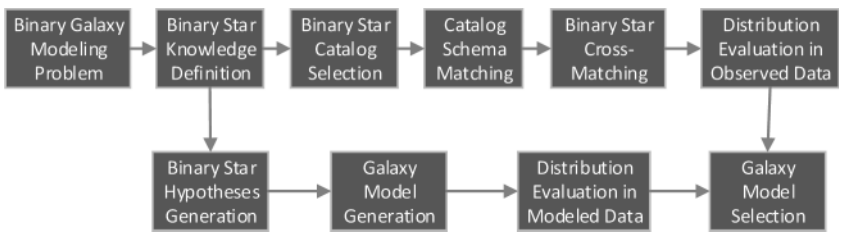


Figure 5. The steps of solving the problem of binary star Galaxy creation.

- The step of creating a problem requirement model can lead to the detection of common sub-steps of the problems.

- At the knowledge specification step, the requirement model can refer to the domain ontologies and published data schemas. Some new concepts should be defined and can be published to enhance the domain knowledge if the community confirms their commitment.
- The ontologies can be used to search for relevant registered catalogs.
- The step of data model integration can be skipped since the catalogs use the same principles of record representations. It can be verified that the empty activity can just transfer the schemas to the next activity.
- At the schema matching step, the results of integrating the binary star catalogs can be retrieved and reused.
- At the entity resolution step, the algorithm for cross-matching binary stars can be partially reused for two-component stars only.
- At the operationalization step, methods for compiling data on star systems, hypotheses and methods for generating models, and methods for evaluation of stellar parameter distributions in the galaxy should be developed since there were no relevant existing methods found. The developed specifications can be published.
- At the experimenting step, experiments should be performed to generate different models and compare stellar parameter distributions to observed ones. The results of the research should be published.

Such refinements of the canonical workflow become possible as a result of referring the steps to concepts of domain ontologies and searching for relevant implementations for them in the library of contextual steps and the published results of previous research.

The considered examples solve specific problems in astronomy, however, most research domains have the necessary resources for proposed approaches. Similar principles of distinguishing types of research objects and observable and computable parameters can be found in almost any area, including astronomy and astrophysics, materials science, biomedicine, earth sciences, social science, and others. This is reflected in existing domain models, in which the same basic principles of research are significantly duplicated. The most general knowledge in these areas can be reduced to common ontologies, data and metadata schemas, standard methods, and processes.

7. ISSUES OF IMPLEMENTATION WITHIN THE CANONICAL WORKFLOW FRAMEWORK

First of all, the proposed ontological approach can be implemented if the chosen implementations of CWFR support semantic annotations in terms of ontologies. For example, elements of workflows can be supported with expressions of concepts so that individuals of ontological concepts can store their identifiers to refer to relevant activities, places (inputs and outputs), and types.

The following references can be used:

- Canonical steps refer to concepts of the ontology determining the meaning of a step in the research lifecycle.

For example, the entity resolution step of data resource integration can refer to the concept "entity resolution" in an ontology of data management and data-driven research.

- Contextual packages of steps also refer to the ontology of the context domain.

For example, the package implementation of the entity resolution step in astronomy can refer to the concept of "cross-matching", which is restricted to be applied only to data on astronomical objects and can be a subconcept of "entity resolution".

- The implementations of certain activities specific to the research domain refer to ontological expressions that reflect their precise semantics in the domain.

For example, an implementation of the activity of binary star cross-matching can be a very complex problem requiring research and making a separate sub-workflow. Both the activity and the sub-workflow can be semantically defined by the expression describing the cross-matching of binary stars only. Such an expression defines a concept that does not belong directly to the ontology but precisely defines the semantics of the described activity.

- The places of workflow modeling patterns correspond to the inputs and outputs of the canonical steps. They should be described by concepts of the workflow modeling pattern ontology.
- The places of contextual package steps and workflow implementations should additionally be described in terms of domain ontologies to constrain data types. They also define the types of input attributes of activities and the types of digital objects created as a result of the activities.
- The library of contextual packages can contain general steps related to domain types. Besides this, the activities of all published workflow implementations for solving specific research problems should be classified in terms of corresponding domain ontologies.

The search for relevant reusable workflow implementations can be organized based on queries to the ontology. The requirements of the query actually should be defined as a concept by which referred workflows and their fragments are classified. The query can simultaneously use constraints of several ontologies: for example requirements from the point of view of the domain, defined workflow element kinds, the provenance of the workflow, and others. Thus, the IDs of all relevant fragments are retrieved by the query for reuse.

During the reuse of existing implemented activities, steps, or workflows, the digital objects are supplemented with metadata in terms of the provenance ontology to track what data were processed, where they came from, which workflows processed the data, which agents executed the workflows and processed the data, when the data was processed, how long it is relevant, and so on.

As for the selection of possible workflow management systems for CWFR implementation, different workflow environments may include semantic descriptions of both the resources used in processes and the processes themselves. RO-Crate supports semantic metadata based on types schema.org and similar dictionaries that can be linked to Research Object containers and resources in them [14]. UIMA defines

types and features identified during data analysis that can be associated with RDF resources. It allows relating metadata with resources and workflows through resource specifiers for organizing registries of reusable resources [15]. There are investigations related to the ontological description of the OPC UA [16] specifications, which abstractly describes nodes and services for organizing data exchange between them. Jupyter Notebook workflow extensions can be used if any definition of a workflow can be identified to link metadata. These and other workflow management tools can be supplemented with instruments for describing resource semantics and semantic search. What an object must have to be described with semantic metadata is an identifier. Ontology-supporting instruments can be independent of the workflow management tools being used. They store both domain definitions and metadata that describe resources and refer to the resources through identifiers. If possible, workflow specifications should keep references to the metadata as well. By semantic metadata, relevant resources can be found and reused.

The proposed prototype architecture implementing the approach as a complementing part of possible architectural decisions in CWFR is shown (Figure 6). A service for resource annotation allows describing resources such as source data, programs, workflows, and resulting data are annotated in terms of domain models and some special models. It is necessary to link annotations to corresponding resource identifiers and keep them in the triple store as RDF/OWL specifications. They are considered as a part of specified digital objects, so the links to the annotations as metadata are stored in the digital objects too. Annotations of workflows needed to be implemented can be considered as simple queries for relevant resources by their similar annotations. A query building service is necessary for covering the requirements to necessary resource compositions like workflow fragment structures. The SPARQL endpoint is used to search for relevant resource IDs. The retrieved resources can be considered as candidates to be reused for consistent implementation of workflows.

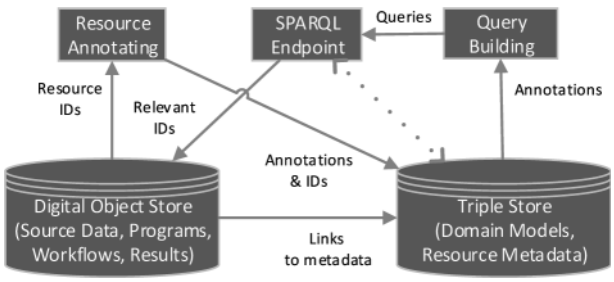


Figure 6. The prototype architecture.

8. FOLLOWING THE PRINCIPLES OF FAIR DATA

The ontological approach to providing the reuse of canonical workflow steps and implemented workflows was motivated by the principles of FAIR data.

To provide the findability of workflows and processed data, the workflows and all their components are widely, comprehensively, and formally described with semantic annotation metadata in terms of domain

ontologies and some special ontologies. Workflows are classified using ontological reasoning [17] over those metadata. The publishing workflows consists in defining semantic annotations and registering them by classification. Query expressions in terms of ontological concepts are used to search for relevant workflows and their fragments.

Data accessibility is provided by referring ontological annotations to identifiers of workflows and their components and returning them by queries. The SPARQL [18] access point interface and web protocols could be used for it.

Interoperability of workflows and data in the proposed approach is provided by the use of a common formal knowledge representation model [19] for all specifications including domain models, workflow language models, research step models, provenance models. Automatic logical reasoning for ontologies on semantic annotations allows searching and controlling relevance and meaning of workflows from the point of view of ontologies. Reasoning provides the ability to interpret the meaning of data and resources when solving problems by both a human and a machine. Formal ontologies are a kind of dictionary fulfilling the FAIR data principles the best.

The principle of reusability is provided by supporting domain ontologies that may describe any other domain-specific standards. On the other hand, the use of special ontologies allows defining non-functional requirements for data and workflows such as data provenance, data quality, and other aspects.

9. DISCUSSION

The proposed approach is natural to be applied in a cross-disciplinary space. It is not intended for a specific domain and would be not so effective in a closed environment.

On the one hand, domain communities, which are smaller the more specific the area, work to determine the knowledge of different domains. Such communities use the knowledge of more generalized communities and domains. Most likely, the domains already have a groundwork defining domain concepts, common conceptual schemes, methods, and tools used by the community, as well as known typical processes for solving various tasks. Today, there is a tendency to intensively develop and discuss domain models for the semantic approaches at specialized conferences and workshops devoted to the problems of certain research domains.

On the other hand, in the cross-disciplinary space, there are standards followed by almost any domain community and the most common domains used by everyone in the discipline. All this knowledge can and should be reused in communities.

For example, there are no commonly applied formal ontologies in astronomy, but the Unified Content Descriptors (UCD) [20] standard is a widely used instrument for linking resources with astronomical domain concepts. Common domains of knowledge that are used in all subdomains and in solving almost any

problems are astrometry with the celestial coordinate system, multicolored photometry, classifications of astronomical objects and stars. Almost any task in this domain includes a subtask of cross-matching data about astronomical objects in different catalogs. There are standards of the International Virtual Observatory that are widely used.

Finally, the most general knowledge is interdisciplinary and is shared by everyone involved in research. For example, the ontology and metadata of the data provenance, measurements, and their accuracy, experiments at the physical and the informational levels. Communities of any domain can have common or own standards for those issues. Most research areas distinguish research objects of various types, sets of observable parameters that can be measured by observation instruments, and sets of parameters that cannot be observed directly but can be estimated based on the observed parameters. The type of the research object may depend on the composition of parameter values, or, on the contrary, the type of object limits the possible parameter values. All this knowledge can have common specifications across domains.

Research skills of any domain researcher allow relating resources to the domain concepts, or even defining constraints combining several concepts to express the semantics of resources. If there are some arrangements for semantic modeling in the domain, they can be naturally used to describe resources with them. Domain researchers should be documenting any resources applied by them including source data, schemas, workflows, activities in them, data at the inputs and outputs of activities, result datasets in terms of their native domain concepts. This is often not done because there are no requirements to provide such descriptions. However, data management plans could include such requirements. Having detailed annotations of resources with domain knowledge, a machine can offer the most relevant resources to substitute and implement workflows and to solve domain problems.

Thus, the proposed approach is not so domain-driven but community-driven. To be alive, a community should define its domain, useful commonalities, shared resources in them, and reuse resources available in its domain and a wider context. To implement an interdisciplinary environment, it is necessary to support domain communities and include them in more general communities with access to common knowledge. Each community organizes its part of the resource registry: domain ontologies, schemas, programs and libraries, processes. The semantics of resources should be described in terms of domain concepts to classify them in the domain. To reuse resources, it is necessary to provide a search in the registry by domain concepts, access by retrieved identifiers, support integration, and interoperability of found resources.

10. CONCLUSION

A semantic approach to workflow search, implementation, and composition in the context of investigations of canonical research workflows has been represented. Specifying workflow model semantics has been proposed for interoperability of possible different workflow systems and execution of workflow fragments in different management systems. Specifying domain semantics of workflows using ontologies can be applied for organizing workflow classification and relevant workflow search, correct pattern implementation, and workflow fragment integration and reuse. Examples of patterns applicable to any research domain and

examples of their implementation for solving specific problems in astronomy were described. The principles of application of the presented approach to CWFR concepts and a prototype architecture implementing this approach as a complementing part of possible architectural decisions in CWFR are proposed.

ACKNOWLEDGEMENTS

The work was carried out using the infrastructure of shared research facilities CKP "Informatics" of FRC CSC RAS [21], and supported by the Russian Foundation for Basic Research, grants 19-07-01198, 18-29-22096.

AUTHOR CONTRIBUTIONS

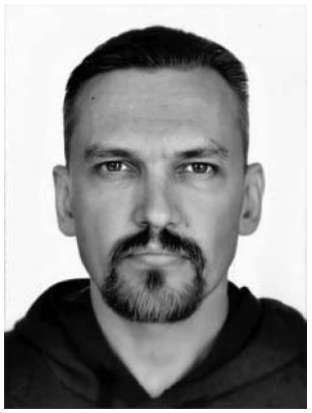
N. Skvortsov (nskv@mail.ru) investigated the ontology-based approach to workflow search, refinement, and composition. S. Stupnikov (sstupnikov@ipiran.ru) proposed the approach to workflow management model integration using workflow control patterns.

REFERENCES

[1] Wilkinson, M., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, Article No. 160018 (2016)
[2] Hardisty, A., Wittenburg, P. (eds). Canonical workflow frameworks for research (CWFR). Available at: https://www.rd-alliance.org/canonical-workflow-frameworks-research-cwfr.
[3] CWFR. OCF (2020). Available at: https://osf.io/2cy86/. Accessed 29 December 2021
[4] Wittenburg, P.: From persistent identifiers to digital objects to make data science more efficient. Data Intelligence 1(1), 6–21 (2019)
[5] Newman, D.: The building and application of a semantic platform for an e-research society. PhD dissertation, University of Southhampton (2011)
[6] Belhajjame, K., et al.: Workflow-centric research objects: A first class citizen in the scholarly discourse. In: Proceedings of the 2nd Workshop on Semantic Publishing (SePublica2012), pp. 1–12 (2012)
[7] Skvortsov, N.A., et al.: Metadata model for semantic search for rule-based workflow implementations. Systems and Means of Informatics 24(4), 4–28 (2014)
[8] Belhajjame, K., et al.: PROV-O: The PROV ontology. W3C Recommendation. W3C (2013). Available at: https://www.w3.org/TR/prov-o. Accessed 29 December 2021
[9] Russell, N., van der Aalst, W.M.P., ter Hofstede, A.H.M.: Workflow patterns: The definitive guide. MIT Press, Cambridge (2016)
[10] Cilia, N.D., Scarpato, N., Romano, M.: A semantic approach to reachability matrix computation. In: The Tenth Conference on Semantic Technology for Intelligence, Defense, and Security (STIDS), pp. 91–94 (2015)
[11] Skvortsov, N.A., Stupnikov, S.A.: Managing data-intensive research problem-solving lifecycle. In: Data Analytics and Management in Data Intensive Domains (DAMDID 2020), pp. 3–18 (2021)
[12] Skvortsov, N.A.: Conceptual model reuse for problem solving in subject domains. In: Modelling to Program (M2P 2020), pp. 191–211 (2021)

[13] Skvortsov, N.A., et al.: Matching and verification of multiple stellar systems in the identification list of binaries. In: Kalinichenko, L., et al (eds.) Data Analytics and Management in Data Intensive Domains, (DAMDID/RCDL 2017), pp. 102–112. Springer, Berlin (2018)
[14] Sefton, P., et al.: RO-crate metadata specification. ResearchObject.org, 2021. Available at: https://zenodo.org/record/5841615#.YgmksPnj7oY. Accessed 29 December 2021
[15] Apache UIMA Java SDK 3.2.0 user-level API documentation. Apache UIMA, 2021. Available at: https://uima.apache.org/d/uimaj-current/apidocs/. Accessed 29 December 2021
[16] Schiekofer, R., et al.: A formal mapping between OPC UA and the semantic Web. In: The 17th International Conference on Industrial Informatics (INDIN), pp. 33–40 (2019)
[17] Baader, F., et al.: Introduction to description logic. Cambridge University Press, Cambridge (2017)
[18] SPARQL query language for RDF. W3C (2008). Available at: http://www.w3.org/TR/rdf-sparql-query/. Accessed 29 December 2021
[19] OWL 2 Web ontology language document overview (Second Edition). W3C (2011). Available at: http://www.w3.org/TR/owl-overview/. Accessed 29 December 2021
[20] An IVOA standard for unified content descriptors. Version 1.1. IVOA, 2005. Available at: https://www.ivoa.net/documents/latest/UCD.html. Accessed 29 December 2021
[21] Regulations of CKP “Informatics”. Available at: http://www.frccsc.ru/ckp. Accessed 29 December 2021

AUTHOR BIOGRAPHY



Nikolay Skvortsov has been affiliated with the Institute of Informatics Problems, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, Russia. His general research interests are ontological and conceptual modeling of research domains and data semantic interoperability issues. In recent years Nikolay Skvortsov investigates requirements for the reuse of data, research methods, and processes in research communities primarily using examples of problem development and solving in astronomical research domains. His research areas also include social network analysis,information security,and heterogeneous data integration. The courses on conceptual modeling of research domains and data analysis in social environments were developed for the master program on Big data at the Department of Computational Mathematics and Cybernetics, Moscow State University.
ORCID: 0000-0003-3207-4955

Sergey Stupnikov graduated from the Lomonosov Moscow State University in 2000 with the specialist degree in Mathematics and Applied Mathematics. He was awarded his Ph.D. degree in Theoretical Computer Science in 2006 at the Institute of Informatics Problems, Russian Academy of Sciences. His current positions are head of department and lead research scientist at the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; acting head of master program entitled "Big data: infrastructures and methods for problem solving" and lecturer at the Department of Computational Mathematics and Cybernetics, Lomonosov Moscow State University. His research areas include heterogeneous data integration; data models and specification languages: object, ontological, process and workflow models, graph-based and array-based data models; data model integration, mapping, transformation, formal semantics and verification. He is author or co-author of 90 publications in international and national journals and conference proceedings. He has been Chair of Moscow ACM SIGMOD Chapter since August 2018, Deputy Chair of the DAMDID/RCDL Conference Coordinating Committee since November 2018, Member of the ADBIS Conference Steering Committee, PC co-chair of several DAMDID/RCDL conferences, and PC member of ADBIS, DAMDID/RCDL, DATA, I-ESA (2012-2018), and SEIM conferences.
ORCID: 0000-0003-4720-8215